Publication VII

Emil Eirola, Amaury Lendasse, Francesco Corona, and Michel Verleysen. The Delta Test: The 1-NN Estimator as a Feature Selection Criterion. In *The 2014 International Joint Conference on Neural Networks (IJCNN 2014)*, pages 4214–4222, July 2014.

The Delta Test: The 1-NN Estimator as a Feature Selection Criterion

Emil Eirola, Amaury Lendasse, Francesco Corona, and Michel Verleysen

***Abstract*— Feature selection is essential in many machine learning problem, but it is often not clear on which grounds variables should be included or excluded. This paper shows that the mean squared leave-one-out error of the first-nearest-neighbour estimator is effective as a cost function when selecting input variables for regression tasks. A theoretical analysis of the estimator's properties is presented to support its use for feature selection. An experimental comparison to alternative selection criteria (including mutual information, least angle regression, and the RReliefF algorithm) demonstrates reliable performance on several regression tasks.**

I. Introduction

With evolving technology and the continuing development of more efficient data mechanisms, the size and complexity of interesting regression modelling tasks has grown considerably. The number of input variables which can be measured might be large, and it may be difficult to recognise which variables are important for a given task. Occasionally variables are irrelevant for the output, and at other times they contain redundant information already available in other variables. Hence the concept of *feature selection* has become increasingly important [1]. Being able to discard the unnecessary ones is beneficial for the performance and stability of the model. Identifying the most essential variables also provides a better understanding of the problem and interpretability of the data. In a sense, feature selection improves the signal-to-noise ratio by getting rid of some of the noisy components.

For linear problems, the issue can be solved by simple covariance or cross-correlation methods [2], [1]. In the case of nonlinear problems, the situation is less straightforward, and there are many specialised methods to pick from, often requiring parameters which are nontrivial to tune. Some methods only rank variables, but cannot tell you how many to choose. The ranking of variables also might fail to identify situations where some variables are useful only in combination with others.

In more general feature selection procedures, some optimisation scheme is applied to search the space of subsets of variables in order to minimise a given cost function. However, there are few generally applicable cost functions, and using specialised or model-specific choices is often inconvenient. In this paper, the Delta test (the mean squared leave-one-out error of the first-nearest-neighbour estimator) is studied and shown to be generally efficient as a cost functions, both in retaining all the important variables as well as excluding irrelevant and redundant variables. The Delta test has no parameters to tune, making it reliable and easy to use. Previously, the Delta test has been used for noise variance estimation, and in a sense represents the lowest attainable mean squared error by quantifying the extent of the "random" part of the data [3]. The idea here then is to use the noise variance estimate given by the Delta test to evaluate a subset—or selection—of variables.

We initially introduced the method in [4], and it has been used with success in several cases [5], [6], [7], [8], [9], [10], [11], [12]. This paper presents further theoretical justification to explain why the method works as well as it does. Experimental evidence comparing to alternative methods is also provided. Choosing an efficient search scheme for high-dimensional tasks is mostly left as a practical matter of implementation, and several papers specifically about optimising the Delta test have also been published in the literature [13], [14], [15], [16], [17], [18], [19].

A similar method for variable selection has previously been proposed in [20], using a weighted k-NN estimator, but has two parameters to optimise (k and a decay factor β).

The sequel of this article is organised as follows: Section II reviews the concepts of variable selection and noise variance estimators and describes the Delta test algorithm. Section III provides the main contribution which is the theoretical justification for the methodology, also discussing why the Delta test cannot simply be replaced by a more accurate noise variance estimator. In Section IV, the use of the method from a practical point of view is considered including how the method can be used with different search schemes. Sections V and VI include several experiments which illustrate the behaviour of the method in different situations and comparisons to other methods.

II. Problem Definition: Variable Selection

In modern modelling problems it is not uncommon to have an overwhelming number of input variables. Many of them may turn out to be irrelevant for the task at hand, but without external information it is often difficult to identify these variables. *Variable selection* (also known as *feature extraction*, *subset selection*, or *attribute selection* [20]) is

Emil Eirola and Francesco Corona are with the Department of Information and Computer Science, Aalto University, Finland

Amaury Lendasse is with the Department of Information and Computer Science, Aalto University, Finland, Arcada University of Applied Sciences, Helsinki, Finland, IKERBASQUE, Basque Foundation for Science, Spain, and the Department of Mechanical and Industrial Engineering, The University of Iowa, USA.

Michel Verleysen is with the Machine Learning Group, Université catholique de Louvain, Louvain-la-Neuve, Belgium, and the SAMOS team, Université Paris I Panthéon-Sorbonne, France

the process of automating this task of choosing the most representative subset of variables for some modelling task.

Variable selection is a special case of dimensionality reduction. It can be used to simplify models by refining the data through discarding insignificant variables. Since many regression models and other popular data analysis algorithms suffer from the so-called *curse of dimensionality* [21] to some degree, it is necessary to perform some kind of dimensionality reduction to facilitate their effective use.

In contrast to general dimensional reduction techniques, variable selection provides additional value by distinctly specifying which variables are important and which are not [1]. This leads to a better insight into the relationship between the inputs and outputs, and assigns interpretability to the input variables. In cases where the user has control over some inputs, variable selection emphasises which variables to focus on and which are likely to be less relevant. Furthermore, discarding the less important inputs may result in cost savings in cases where measuring certain properties could be expensive (such as chemical properties of a sample).

A. The Delta Test

The Delta test is traditionally considered a method for residual noise variance estimation. In the kind of regression tasks considered here, the data consist of M input points $\{\boldsymbol{x}_i\}_{i=1}^M$ and associated scalar outputs $\{y_i\}_{i=1}^M$ [22]. The assumption is that there is a functional dependence between them with an additive noise term: $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$.

The function f is often assumed to be smooth, or at least continuous, and the additive noise terms ε_i are i.i.d. with zero mean and finite variance. Noise variance estimation is the study of how to find an a priori estimate for $\mathrm{Var}(\varepsilon)$ given some data without considering any specifics of the shape of f. Having a reliable estimate of the amount of noise is useful for model structure selection and determining when a model may be overfitting.

The original formulation [23] of the Delta test was based on the concept of variable-sized neighbourhoods, but an alternative formulation [24] with a first-nearest-neighbour (NN) approach has later surfaced. In this treatment, specifically this 1-NN formulation will be used as it is entirely non-parametric, conceptually simple, and computationally efficient. The Delta test could be interpreted as an extension of the Rice variance estimator [25] to multivariate data.

The NN of a point is defined as the unique point in a data set which minimises a distance metric to that point:

$$N(i) := \arg\min_{j \neq i} \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|^2 . \tag{1}$$

It may occur that the nearest neighbour is not unique, and in that case it is sufficient to randomly pick one from the set of nearest neighbours. In this context, the Euclidean distance is used, but in some cases it may be justified to use other metrics to get better results if some input variables are known to have specific characteristics that the Euclidean metric fails to account for appropriately. Knowing if other metrics are more appropriate generally requires external knowledge about the source or behaviour of the data. For instance, data representing class labels are best handled by the discrete metric, and "time-of-day" or "time-of-year"-type variables by taking into account their cyclic behaviour.

The Delta test, initially introduced in [23] and further developed in [24], is usually written as

$$\delta = \frac{1}{2M} \sum_{i=1}^{M} \left(y_i - y_{N(i)} \right)^2 , \tag{2}$$

i.e., the differences in the outputs associated with neighbouring (in the input space) points are considered. This is a well-known estimator of $\mathrm{Var}(\varepsilon)$ and it has been shown—e.g., in [26]—that the estimate converges to the true value of the noise variance in the limit $M \to \infty$. Although it is not considered to be the most accurate noise estimator, its advantages include reliability, simplicity, and computational efficiency [22]. The method appears not to be particularly sensitive to mild violations of the assumptions made about the data, such as independence and distributions of the noise terms.

B. The Delta Test for Variable Selection

The Delta test was originally intended to be used for estimating residual variance. Following [7], [9] this paper examines a different use: to use it as a cost function for variable selection by choosing that selection of variables which minimises the Delta test. Each subset of variables can be mapped to a value of the estimator by evaluating the expression in eq. (2) so that the nearest neighbours $N(i)$ are determined according to the distance in the subspace spanned by the subset of variables. The sequel of this paper intends to show that choosing the subset which provides the smallest value for the Delta test constitutes an effective variable selection procedure for regression modelling.

An exhaustive search over the $2^d - 1$ non-empty subsets of d variables is a possibility, but there are more efficient approximate search schemes as discussed in Section IV.

III. Theoretical Considerations

In this section, a theoretical treatment is provided to support the claim that the Delta test is able to identify the best subset of input variables for modelling with high probability under appropriate conditions. As the purpose of the Delta test is to deal with noisy data, it is impossible to formulate a mathematically solid statement showing that the Delta test could always choose the perfect variables, due to the random effects of the noise. Hence the assertions presented here consider the expectation of the Delta test, and show that the expectation is minimised for the best selection of variables for a finite number M of data points.

A. Analysis of the Delta Test

Some assumptions concerning the distribution of the data are required in order for the results to hold true. These continuity assumptions detailed below are designed to be similar to and compatible with the assumptions many popular non-linear modelling techniques make about the data. This

enhances the usability of the Delta test as a preprocessing step for practically any non-linear regression task.

Assume a set $\{X_i\}_{i=1}^M$ of random variables which are independently and identically distributed according to some probability density function $p(\boldsymbol{x})$ for $1 \leq i \leq M$. Here $p(\boldsymbol{x})$ is a continuous probability density on some open, bounded $C \subset \mathbb{R}^d$ and $p(\boldsymbol{x}) > 0$ for $\boldsymbol{x} \in C$.

Let $f : C \to \mathbb{R}$ be differentiable and the random variables $Y_i = f(X_i) + \varepsilon_i$, where ε_i are independently distributed according to some distribution with mean 0 and finite variance σ^2. Denote by $\boldsymbol{x}_i$ a realisation of X_i and by y_i a realisation of Y_i. The component k of a vector $\boldsymbol{x}$ is denoted by $x^{(k)}$.

Let $I = \{1, \ldots, d\}$ denote the full set of input variables, and consider subsets $\tilde{I}$ of I corresponding to possible selections of input variables. If $\tilde{I} \subset I$ then $\tilde{\boldsymbol{x}}_i = \left(x_i^{(\tilde{I}_1)}, x_i^{(\tilde{I}_2)}, \ldots\right)$ is the projection of each point to the subspace corresponding to the selected variables. Similarly $\tilde{x}'_i$ includes the components not in $\tilde{I}$ and is the projection onto the subspace corresponding to $I \setminus \tilde{I}$.

Define $\tilde{f}(\tilde{\boldsymbol{x}}_i)$ as the conditional expectation of the output with partial information:

$$\tilde{f}(\tilde{\boldsymbol{x}}_i) = \mathrm{E}\big[Y_i \mid \tilde{X}_i = \tilde{\boldsymbol{x}}\big]\,. \tag{3}$$

Now $\tilde{f}(\tilde{\boldsymbol{x}}_i)$ is the best possible approximation of y_i (in terms of mean squared error) when using only the variables in $\tilde{I}$. In particular, if $\tilde{f}(\tilde{\boldsymbol{x}}) = f(\boldsymbol{x})$ for all $\boldsymbol{x} \in C$, it can be said that the variables in $\tilde{I}$ hold all the information for fully determining $f(\boldsymbol{x})$.

Using $\mathcal{P}\big(I\big)$ to denote the power set of I, define the Delta test $\delta : \mathcal{P}\big(I\big) \to \mathbb{R}$ as

$$\delta(\tilde{I}) := \tfrac{1}{2M} \sum_{i=1}^{M} \big(y_i - y_{N_{\tilde{I}}(i)}\big)^2 \tag{4}$$

where

$$N_{\tilde{I}}(i) := \arg\min_{j \neq i} \big\|\boldsymbol{x}_i - \boldsymbol{x}_j\big\|_{\tilde{I}}^2\,, \tag{5}$$

and the semi-norm

$$\big\|\boldsymbol{x}_i - \boldsymbol{x}_j\big\|_{\tilde{I}}^2 := \sum_{k \in \tilde{I}} \big(x_i^{(k)} - x_j^{(k)}\big)^2\,. \tag{6}$$

The argument is split into two lemmas which together imply the main result. Even if the results specifically deal with the expectation of the Delta test, they provide insight into how the method behaves even for a single realisation of data. In the following, the phrase "for any sufficiently large M ..." should be interpreted as "$\exists M_0 < \infty$ s.t. $\forall M \geq M_0$ ...". The symbol $\#$ is used for the cardinality of a set.

Lemma 1: If $\tilde{I}$ is such that $\exists \boldsymbol{x}_0 \in C$ for which $\tilde{f}(\tilde{\boldsymbol{x}}_0) \neq f(\boldsymbol{x}_0)$ (i.e., the variables in $\tilde{I}$ are not sufficient to explain f) then for any sufficiently large M

$$\mathrm{E}\big[\delta\big(\tilde{I}\big)\big] > \mathrm{E}\big[\delta\big(I\big)\big]\,.$$

Proof: According to [26], the estimate for an incomplete selection of variables converges to the residual noise

$$\lim_{M\to\infty} \mathrm{E}\big[\delta\big(\tilde{I}\big)\big] = \mathrm{E}\big[\big(Y_i - \tilde{f}(\tilde{X}_i)\big)^2\big] \tag{7}$$

and, correspondingly,

$$\lim_{M\to\infty} \mathrm{E}\big[\delta\big(I\big)\big] = \mathrm{E}\big[\big(Y_i - f(X_i)\big)^2\big]\,. \tag{8}$$

Furthermore,

$$\begin{aligned}\mathrm{E}\big[\big(Y_i - \tilde{f}(\tilde{X}_i)\big)^2\big] &= \mathrm{E}\big[\big(Y_i - f(X_i) + f(X_i) - \tilde{f}(\tilde{X}_i)\big)^2\big] \\ &= \mathrm{E}\big[\big(Y_i - f(X_i)\big)^2\big] + \mathrm{E}\big[\big(f(X_i) - \tilde{f}(\tilde{X}_i)\big)^2\big]\end{aligned}$$

since the cross terms cancel by the independence of the noise. Now

$$\mathrm{E}\big[\big(f(X_i) - \tilde{f}(\tilde{X}_i)\big)^2\big] = \int_C \big(f(\boldsymbol{x}) - \tilde{f}(\tilde{\boldsymbol{x}})\big)^2 p(\boldsymbol{x})\, d\boldsymbol{x} > 0 \tag{9}$$

where the integral is positive because the continuities of f, $\tilde{f}$, and p together imply that there is an open subset of C around $\boldsymbol{x}_0$ where $\tilde{f}(\tilde{\boldsymbol{x}}) \neq f(\boldsymbol{x})$ and $p(\boldsymbol{x}) > 0$. Since the term is independent of M, the difference

$$\lim_{M\to\infty} \mathrm{E}\big[\delta\big(\tilde{I}\big)\big] - \mathrm{E}\big[\delta\big(I\big)\big] = \int_C \big(f(\boldsymbol{x}) - \tilde{f}(\tilde{\boldsymbol{x}})\big)^2 p(\boldsymbol{x})\, d\boldsymbol{x} > 0$$

is positive even in the limit $M \to \infty$, implying there exists an M_0 such that the expression is positive for any $M \geq M_0$. Hence, for sufficiently large M, the first term is larger, proving the lemma. ∎

Lemma 2: If $\tilde{I}$ and $\hat{I}$ are such that $\forall i : \tilde{f}(\tilde{\boldsymbol{x}}_i) = \hat{f}(\hat{\boldsymbol{x}}_i) = f(\boldsymbol{x}_i)$—i.e., they are both sufficient to explain f—and $\#\tilde{I} < \#\hat{I}$, then for any finite and sufficiently large M

$$\mathrm{E}\big[\delta\big(\tilde{I}\big)\big] < \mathrm{E}\big[\delta\big(\hat{I}\big)\big]\,.$$

Proof:

$$\begin{aligned}\mathrm{E}\big[\delta\big(\tilde{I}\big)\big] &= \tfrac{1}{2}\,\mathrm{E}\big[\big(Y_i - Y_{N_{\tilde{I}}(i)}\big)^2\big] \\ &= \tfrac{1}{2}\,\mathrm{E}\big[\big(f(X_i) + \varepsilon_i - f(X_{N_{\tilde{I}}(i)}) - \varepsilon_{N_{\tilde{I}}(i)}\big)^2\big] \\ &= \tfrac{1}{2}\,\mathrm{E}\big[\varepsilon_i^2 + \varepsilon_{N_{\tilde{I}}(i)}^2\big] + \tfrac{1}{2}\,\mathrm{E}\big[\big(f(X_i) - f(X_{N_{\tilde{I}}(i)})\big)^2\big]\end{aligned}$$

as the noise terms are independent, and further, as $f = \tilde{f}$,

$$= \sigma^2 + \tfrac{1}{2}\,\mathrm{E}\big[\big(\tilde{f}(\tilde{X}_i) - \tilde{f}(\tilde{X}_{N_{\tilde{I}}(i)})\big)^2\big]$$

where the first term is obviously identical for $\tilde{I}$ and $\hat{I}$. According to [27] the second term is of order $M^{-2/\#\tilde{I}}$. So, for a sufficiently large M, this will be the dominating term, implying that a smaller selection produces a smaller Delta test estimate, proving the lemma. ∎

Theorem 1: For any finite and sufficiently large M, the expectation of the Delta test is minimised by the smallest subset of I which can fully explain f on C.

Proof: Provided the number of points is sufficiently large, by Lemma 1 the minimising selection must be able to fully explain f, and by Lemma 2 it must be the smallest such selection. ∎

It is shown in [26] that the variance of the Delta test converges to 0 with increasing M. As the expectation of the Delta Test under the above assumptions is strictly minimised by the "best" selection, this means that the probability of the method choosing this selection generally increases by increasing the number of available samples.

B. On other noise estimators

While the motivation for using the Delta test for variable selection appears to originate from its role as a noise variance estimator, it can not simply be replaced by a more accurate estimator—such as those presented by [22], [24], [28], [29]. On some level, it is intuitively sensible to optimise a model by "minimising the noise", but it is far from obvious whether the proposed scheme is justified beyond that. In Section III-A, it is shown to hold for the Delta test, but it is worth investigating if other noise estimators can be used in this way.

As in the proof of Lemma 2 previously, the expectation of the Delta Test can be expanded as

$$\mathrm{E}\big[\delta(\tilde{I})\big] = \sigma^2 + \tfrac{1}{2}\,\mathrm{E}\big[\big(f(X_i) - f(X_{N_{\tilde{I}}(i)})\big)^2\big]\,. \qquad (10)$$

Here it is clearly seen that unless f is constant, the Delta Test has a positive bias in its role as a noise variance estimator (for a finite M; in the limit $M \to \infty$, the bias naturally converges to 0). This makes the estimator relatively poor for estimating noise compared to more sophisticated, less biased alternatives. However, the proposed method works for variable selection effectively by exploiting this bias. All noise estimators should be able to identify the important variables (since excluding one would inflate the noise estimate) but the Delta test has the unique ability to also prune unnecessary variables. This is because improved noise estimators are generally designed to be unbiased, so they do not have the property that the bias increases with the number of selected variables.

IV. Practical Considerations

When using the Delta test, it is important to consider standardising the data beforehand. In particular, the variance of the input variables needs to be of the same order for the method to be effective. Otherwise, the variables with larger variance will have an artificially inflated significance in the selection. For this reason, standardising the variables to unit variance is generally advisable.

The naïve implementation of the Delta test is to separately find the nearest neighbour of each point, leading to a complexity of $O(M^2)$ per evaluation. However, this can be improved by using k-d trees [30] to determine the nearest neighbours, leading to an expected complexity of $M \log M$, even if the worst case complexity is still M^2. On the other hand, performing an exhaustive search over the space of all possible selections requires $2^d - 1$ evaluations. Our rule of thumb is that on a conventional, reasonably modern, desktop computer an exhaustive search takes a few seconds for a data set with $M = 1000$ points and $d = 10$ variables. The exhaustive search is then practical up to 10 or at most 20 variables. However, many interesting problems are much larger than this.

Due to the nature of the noise variance estimator, it is often not necessary to find the selection providing the global minimum test value. Rather, a pragmatic approach is that reducing the estimate generally results in a better selection. Based on this notion it is beneficial to use different heuristics for searching the space of all possible selections:

- The *sequential forward selection* [7] method, which starts from the empty selection and proceeds by sequentially adding that variable which results in the best improvement of the Delta test. Similarly, the *sequential backward elimination* method starts from the full selection and iteratively removes variables. Each method requires at most d evaluations of the Delta test.
- The *forward-backward* (or *stepwise*) search [7], which if started from an empty initialisation is like the forward search, but in addition to adding variables, it also considers the option of removing each of the previously selected variables, and makes the change which improves the target metric the most. This addition/removal of single variables is continued until convergence. The search can also be started from the full selection, or any number of random initialisation, to more extensively explore the search space. The method appears to converge in $O(d)$ steps, requiring $O(d^2)$ evaluations, and has been found to often be effective.
- *Tabu* search [31], which is similar to the forward-backward search, but with additional conditions allowing it to efficiently get out of local minima, leading to better results. This search methodology has been successfully applied to optimising the Delta test for variable selection in [16], [13].

A general overview of search strategies for feature selection by filter methods is found in [1, Chapter 4].

As the Delta test is an estimate of the residual noise, it represents the lowest generalisation error that a model is expected to be able to reach [3]. Alternatively, it can be seen as the lowest possible training error without resorting to overfitting. In fact, as the Delta test has a bias which is always slightly positive, it can be considered a safe choice to train a model until its training error matches the Delta test estimate. Consequently, it is often useful to store the final value of the Delta test in addition to the set of selected variables when using the method.

Another interpretation of the Delta test is that it is half of the leave-one-out error of the 1-NN regression model. This provides another useful metric to compare to when performing model structure selection, since any sophisticated model should be able to perform better than the simple 1-NN model. Essentially, if $\tilde{I}$ is the set of variables that are selected, and $\delta(\tilde{I})$ is the respective value of the Delta test, the leave-one-out or generalisation error of a good non-linear model using the variables $\tilde{I}$ should be between $\delta(\tilde{I})$ and $2\delta(\tilde{I})$, and preferably close to the lower limit.

V. Synthetic Experiment

To illustrate the effectiveness of the procedure, an artificial experiment is conducted to compare the Delta test (DT) to variable selection by *mutual information* (MI). A synthetic experiment is appropriate as it allows repeatable instances of identical setups, and illustrates how the accuracy of the methods improves with increasing sample sizes.

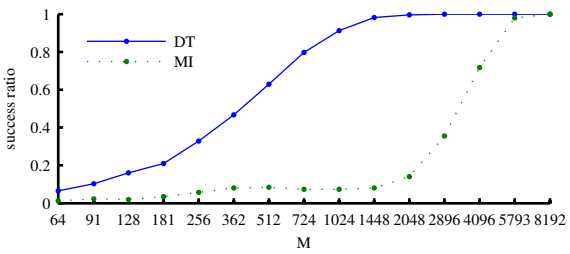

Fig. 1. The convergence of choosing the correct input selection. The vertical axis represents the ratio of cases where each method correctly identified the optimal selection from a total of 1000 tests for each point. Note the logarithmic scale for M.

Mutual information is a quantity which measures the mutual dependencies between two random variables, and the extent to which they can be used to explain each other. Maximising the mutual information has been used as a criterion for variable selection in [32], [7]. The method of [33]—which is also based on nearest neighbours—is used for estimating the mutual information, using the recommended value of $k = 6$ for the number of neighbours.

For this synthetic test, a very non-linear function is intentionally chosen:

$$f(\boldsymbol{x}) = \cos(2\pi x^{(1)}) \cos(4\pi x^{(2)}) \exp(x^{(2)}) \exp(2x^{(3)}) \tag{11}$$

with x distributed uniformly on the unit cube $[0, 1]^6 \subset \mathbb{R}^6$. Obviously, the optimal selection of variables for training a model is $I = \{1, 2, 3\}$. To make the task challenging, the signal-to-noise ratio of the data is made to be 1:1 by choosing the variance of the noise to be equal to the variance of $f(\boldsymbol{x})$:

$$\mathrm{Var}\big[\varepsilon\big] = \mathrm{Var}\big[f(\boldsymbol{x})\big] \approx 10.77.$$

The estimators were given all $2^6 - 1$ non-empty input selections, and the one which minimises the DT or maximises the MI is chosen. The results are presented in Figure 1, where the vertical axis represents the fraction of cases where the correct selection was chosen. The experiment was performed as a Monte Carlo simulation with 1000 repetitions for each value of the data set size M.

It is clear that with increasing data size, the Delta test is eventually able to reliably choose the correct selection, as the curve tends towards 1. The necessary size of about 1000 points in this case might seem high, but recall that the situation was deliberately chosen to be problematic with the high amount of noise. The mutual information method is less successful. Although the success rate does increase with M, the accuracy is much lower for smaller values of M when compared to the Delta test. The method also requires a significantly larger number of points to converge to 1.

VI. Real World Data

In this section, the Delta test is benchmarked on six benchmark datasets consisting of real-world measurements. All the features including the target output are standardised to zero mean and unit variance as a pre-processing step. The Delta test method is compared to variable selection by mutual information [32], variable ranking by least angle regression (LARS) [34], and the RReliefF algorithm [35]. The resulting selections are evaluated by training a least squares support vector machine (LS-SVM) [36] non-linear model using LS-SVMlab v1.8 and calculating the leave-one-out (LOO) mean squared error (MSE). This provides a fair and unbiased method to compare the performance of each method for regression modelling.

LARS is a non-parametric method to rank variables [34]. It provides an exact solution for the optimal variables for the L_1 constrained ordinary least squares regression, but is also commonly used as a feature selection filter for non-linear regression problems.

The Relief algorithm [37] is another popular method for feature selection for classification problems, and has been extended to regression in the RReliefF algorithm [35]. The output of the algorithm is a set of weights for the input variables, which can be converted to a ranking of the variables in order of importance. We apply the method with $k = 10$ nearest neighbours, and a distance scaling factor of $\sigma = 50$.

The LS-SVM is an adaptation of support vector machines with two parameters to choose: the width σ of the Gaussian kernel and a regularisation parameter γ. These hyper-parameters are obtained for each model by running the author-supplied function `tunelssvm` which performs a grid search with initial boundaries specified by certain heuristics and was started with the initial values $\sigma^2 = 1$ and $\gamma = 1$. The function uses a series of random cross-validations to tune the parameters, which unfortunately introduces a certain degree of variability in the results. To eliminate discrepancies caused by random effect where the optimisation gets stuck in local minima, all the LS-SVM models were tuned and evaluated 100 times, and the median error value is used. The leave-one-out error of the LS-SVM is chosen as the measure of performance since the LS-SVM provides an efficient and exact method to calculate it [38], and it is a fair measure of the suitability of the selected variables for modelling.

For the data sets in sections VI-A and VI-B, where the number of variables is low, the minimum DT and maximum MI selections are found by an exhaustive search over all non-empty feature subsets. In the remaining sections, the multi-start sequential search strategy of [17] is used to optimise both DT and MI.

As both the LARS and RReliefF methods only provide a ranking, and not a particular subset, the LS-SVM is sequentially evaluated for each number of variables, successively choosing the top-ranked ones until all variables are selected.

A. Boston Housing

The Boston housing data set [39] is a set with 14 attributes for 506 objects, and the modelling task is to predict the value of a house/apartment from the 13 other properties. The variables selected by DT, MI, LARS, and RReliefF, as well as the median LOO-errors of the LS-SVM are all presented in Table I. There are no obviously redundant variables in

TABLE I

THE SELECTED INPUTS AND MEDIAN LOO MSE FOR THE BOSTON HOUSING DATA. DT VALUE IS 0.0710.

	1	2	3	4	5	6	7	8	9	10	11	12	13	MSE
DT	•		•		•	•	•	•	•	•		•	•	**0.0892**
MI						•				•	•		•	0.1342
LARS 1–01													•	0.3236
LARS 1–02						•							•	0.2323
LARS 1–03						•					•		•	0.2037
LARS 1–04						•					•	•	•	0.1909
LARS 1–05				•		•					•	•	•	0.1772
LARS 1–06	•			•		•					•	•	•	0.1544
LARS 1–07	•			•		•		•			•	•	•	0.1435
LARS 1–08	•			•	•	•		•			•	•	•	0.1331
LARS 1–09	•	•		•	•	•		•			•	•	•	0.1360
LARS 1–10	•	•	•	•	•	•		•			•	•	•	0.1290
LARS 1–11	•	•	•	•	•	•		•	•		•	•	•	0.1162
LARS 1–12	•	•	•	•	•	•		•	•	•	•	•	•	0.1047
Relief 1–01						•								0.4257
Relief 1–02						•							•	0.2323
Relief 1–03					•	•							•	0.1901
Relief 1–04					•	•		•					•	0.1739
Relief 1–05				•	•	•		•					•	0.1829
Relief 1–06				•	•	•		•		•			•	0.1569
Relief 1–07				•	•	•		•		•		•	•	0.1379
Relief 1–08		•		•	•	•		•		•		•	•	0.1317
Relief 1–09		•		•	•	•		•	•	•		•	•	0.1150
Relief 1–10		•		•	•	•		•	•	•	•	•	•	0.1066
Relief 1–11		•	•	•	•	•		•	•	•	•	•	•	0.1053
Relief 1–12		•	•	•	•	•	•	•	•	•	•	•	•	0.0953
All	•	•	•	•	•	•	•	•	•	•	•	•	•	0.0926

TABLE II

THE SELECTED INPUTS AND MEDIAN LOO MSE FOR THE SANTA FE A DATA. DT VALUE IS 0.0165.

	1	2	3	4	5	6	7	8	9	10	11	12	MSE
DT	•	•										•	0.0143
MI	•						•	•					0.0811
LARS 1–01								•					0.3743
LARS 1–02							•	•					0.1250
LARS 1–03			•				•	•					0.1044
LARS 1–04	•		•				•	•					0.0200
LARS 1–05	•	•	•				•	•					0.0137
LARS 1–06	•	•	•	•			•	•					0.0138
LARS 1–07	•	•	•	•	•		•	•					0.0140
LARS 1–08	•	•	•	•	•	•	•	•					0.0152
LARS 1–09	•	•	•	•	•	•	•	•	•				0.0157
LARS 1–10	•	•	•	•	•	•	•	•	•	•			0.0198
LARS 1–11	•	•	•	•	•	•	•	•	•	•		•	0.0209
Relief 1–01	•												0.6538
Relief 1–02	•	•											0.0208
Relief 1–03	•	•						•					**0.0086**
Relief 1–04	•	•						•	•				0.0144
Relief 1–05	•	•					•	•	•				0.0151
Relief 1–06	•	•					•	•	•	•			0.0245
Relief 1–07	•	•				•	•	•	•	•			0.0225
Relief 1–08	•	•	•			•	•	•	•	•			0.0177
Relief 1–09	•	•	•			•	•	•	•	•	•		0.0219
Relief 1–10	•	•	•			•	•	•	•	•	•	•	0.0240
Relief 1–11	•	•	•		•	•	•	•	•	•	•	•	0.0241
All	•	•	•	•	•	•	•	•	•	•	•	•	0.0245

the data set, as is evidenced by the constantly decreasing error when successively choosing the variables ranked by LARS and RReliefF. The DT, also, selects all but three of the available variables. Observing the leave-one-out errors still reveals that the selection by the DT provides a solid improvement in the model accuracy compared to the MI or any of the LARS or RReliefF selections.

The final value of the Delta test is $\delta(\tilde{I}) = 0.0710$ here. The resulting median LOO error of 0.0892 falls appropriately between $\delta(\tilde{I})$ and $2\delta(\tilde{I})$, while being close to the lower value, being in line with what is expected in Section IV.

B. Time Series Prediction: Santa Fe A Laser Data

One interesting application where variable selection is often required is in auto-regressive time series prediction, and hence the methods are also tested on a well known time series problem from the Santa Fe Time Series Competition: the laser data known as Santa Fe A [40], [41]. The series contains 1000 samples of intensity data of a a far-infrared-laser in a chaotic state, and the task is to perform one-step-ahead prediction. It has been shown that a regressor size of 12 should suffice to train an efficient model. The variable selection then pertains to which of the delayed regressors (up to a delay of 12) should be used to build the model.

The results are shown in Table II. The best accuracy by a significant margin is obtained by choosing the top three variables as ranked by RReliefF. The Delta test performs decently, leading to a better model than MI, while choosing only three of the regressor variables, and is on par with the model with four of the top ranked RReliefF variables and the best performing selections by LARS.

C. AnthroKids

The AnthroKids data contains anthropological measurements of children in the USA in 1977 [42]. The full original data included a total 122 measurements of 3900 individuals. As that data has several missing values, it has been converted to a regression problem in [8] by assigning the weight to be the target, and retaining 53 variables and 1019 samples without missing values. In addition to physical attributes, the data contains general information about the individuals and the measurement event. As there are several entirely redundant variables, variable selection should prove effective in improving the performance of the model.

The selected variables with resulting LOO errors are presented in Table III. For LARS and RReliefF, the results are shown only for the 20 highest ranked variables, as the addition of any further variables did not notably decrease the LOO error. The variables selected by the Delta test lead to an error on par with the best selections by RReliefF. MI and LARS result in somewhat less accurate models.

D. Triazines

The regression task is to model the activity level of different drugs (triazines) based on their chemical attributes [43], [44]. There are 186 drug samples and 58 features (after discarding two variables which are constant for all samples). The data contains perfectly collinear sets of attributes, leading to a rank-deficient input matrix. As a result, the LARS algorithm only returns 16 variables.

The results are presented in table IV. The features as selected by the Delta test result in the most accurate LS-

TABLE III

THE SELECTED INPUTS AND MEDIAN LOO MSE FOR THE ANTHROKIDS DATA. DT VALUE IS 0.0084.

		MSE
DT	1,2,3,4,17,19,20,21,27,35,36,37,38,39,40,41	**0.0091**
MI	1,35,37,39	0.0130
LARS 1–01	35	0.0473
LARS 1–02	35,39	0.0283
LARS 1–03	21,35,39	0.0273
LARS 1–04	21,35,37,39	0.0189
LARS 1–05	17,21,35,37,39	0.0171
LARS 1–06	2,17,21,35,37,39	0.0140
LARS 1–07	2,3,17,21,35,37,39	0.0140
LARS 1–08	2,3,17,21,35,36,37,39	0.0131
LARS 1–09	2,3,17,19,21,35,36,37,39	0.0131
LARS 1–10	2,3,17,19,21,33,35,36,37,39	0.0124
LARS 1–11	2,3,17,19,20,21,33,35,36,37,39	0.0100
LARS 1–12	2,3,17,19,20,21,33,35,36,37,39,48	0.0098
LARS 1–13	2,3,17,19,20,21,33,35,36,37,39,44,48	0.0098
LARS 1–14	2,3,17,19,20,21,33,35,36,37,39,44,48,49	0.0097
LARS 1–15	2,3,17,19,20,21,33,35,36,37,39,44,48,49,51	0.0096
LARS 1–16	2,3,17,19,20,21,33,35,36,37,39,44,48,49,51,53	0.0097
LARS 1–17	2,3,17,19,20,21,33,35,36,37,39,44,48,49,51,52,53	0.0099
LARS 1–18	2,3,17,19,20,21,33,35,36,37,39,44,46,48,49,51,52,53	0.0099
LARS 1–19	2,3,16,17,19,20,21,33,35,36,37,39,44,46,48,49,51,52,53	0.0100
LARS 1–20	2,3,16,17,19,20,21,33,35,36,37,39,40,44,46,48,49,51,52,53	0.0100
Relief 1–01	38	0.1075
Relief 1–02	36,38	0.0858
Relief 1–03	36,37,38	0.0505
Relief 1–04	19,36,37,38	0.0327
Relief 1–05	3,19,36,37,38	0.0329
Relief 1–06	3,19,36,37,38,39	0.0263
Relief 1–07	3,19,35,36,37,38,39	0.0196
Relief 1–08	3,19,21,35,36,37,38,39	0.0179
Relief 1–09	3,17,19,21,35,36,37,38,39	0.0167
Relief 1–10	3,17,19,21,35,36,37,38,39,52	0.0169
Relief 1–11	2,3,17,19,21,35,36,37,38,39,52	0.0132
Relief 1–12	2,3,4,17,19,21,35,36,37,38,39,52	0.0098
Relief 1–13	2,3,4,17,19,21,33,35,36,37,38,39,52	0.0094
Relief 1–14	2,3,4,17,19,21,33,35,36,37,38,39,41,52	0.0092
Relief 1–15	2,3,4,17,19,21,24,33,35,36,37,38,39,41,52	0.0092
Relief 1–16	1,2,3,4,17,19,21,24,33,35,36,37,38,39,41,52	**0.0091**
Relief 1–17	1,2,3,4,17,18,19,21,24,33,35,36,37,38,39,41,52	0.0092
Relief 1–18	1,2,3,4,17,18,19,21,24,33,35,36,37,38,39,41,52,53	0.0092
Relief 1–19	1,2,3,4,17,18,19,21,24,26,33,35,36,37,38,39,41,52,53	**0.0091**
Relief 1–20	1,2,3,4,17,18,19,21,24,26,30,33,35,36,37,38,39,41,52,53	**0.0091**
All	All 1–53	1.0010

TABLE IV

THE SELECTED INPUTS AND MEDIAN LOO MSE FOR THE TRIAZINES DATA. DT VALUE IS 0.1655.

		MSE
DT	1,3,4,5,8,9,26,31,32,33,34,35,40,44	**0.4901**
MI	4,5,8,9,32,33,35,40,48	0.6009
LARS 1–01	8	0.9372
LARS 1–02	8,10	0.8270
LARS 1–03	8,10,11	0.7988
LARS 1–04	8,10,11,40	0.7909
LARS 1–05	8,10,11,33,40	0.6319
LARS 1–06	8,10,11,33,40,42	0.6502
LARS 1–07	8,10,11,33,40,42,54	0.6508
LARS 1–08	8,10,11,26,33,40,42,54	0.6498
LARS 1–09	8,10,11,26,33,36,40,42,54	0.6669
LARS 1–10	8,10,11,26,33,36,37,40,42,54	0.6526
LARS 1–11	5,8,10,11,26,33,36,37,40,42,54	0.6378
LARS 1–12	1,5,8,10,11,26,33,36,37,40,42,54	0.6483
LARS 1–13	1,5,8,10,11,15,26,33,36,37,40,42,54	0.6563
LARS 1–14	1,5,8,10,11,15,23,26,33,36,37,40,42,54	0.6581
LARS 1–15	1,5,8,10,11,15,23,24,26,33,36,37,40,42,54	0.6561
LARS 1–16	1,5,8,10,11,15,23,24,25,26,33,36,37,40,42,54	0.6579
Relief 1–01	10	0.9413
Relief 1–02	5,10	0.9472
Relief 1–03	4,5,10	0.8784
Relief 1–04	4,5,6,10	0.8236
Relief 1–05	2,4,5,6,10	0.7979
Relief 1–06	2,3,4,5,6,10	0.7928
Relief 1–07	2,3,4,5,6,8,10	0.7951
Relief 1–08	2,3,4,5,6,8,10,31	0.7834
Relief 1–09	2,3,4,5,6,8,10,31,32	0.6405
Relief 1–10	2,3,4,5,6,8,10,31,32,33	0.5218
Relief 1–11	2,3,4,5,6,8,10,31,32,33,36	0.5708
Relief 1–12	2,3,4,5,6,8,10,11,31,32,33,36	0.5544
Relief 1–13	2,3,4,5,6,8,10,11,31,32,33,36,37	0.5770
Relief 1–14	2,3,4,5,6,8,9,10,11,31,32,33,36,37	0.5588
Relief 1–15	2,3,4,5,6,8,9,10,11,21,31,32,33,36,37	0.5558
Relief 1–16	2,3,4,5,6,8,9,10,11,21,31,32,33,36,37,38	0.5291
Relief 1–17	2,3,4,5,6,8,9,10,11,12,21,31,32,33,36,37,38	0.5717
Relief 1–18	1,2,3,4,5,6,8,9,10,11,12,21,31,32,33,36,37,38	0.5704
Relief 1–19	1,2,3,4,5,6,8,9,10,11,12,15,21,31,32,33,36,37,38	0.5705
Relief 1–20	1,2,3,4,5,6,8,9,10,11,12,15,21,31,32,33,36,37,38,40	0.5725
All	All 1–58	1.0054

SVM model, compared to selection by mutual information, or any of the LARS or RReliefF rankings.

E. Wisconsin Breast Cancer

The data includes several measurements of cancer patients, with the goal to predict the recurrence time [44]. There are 194 patients and 32 continuous variables.

The results in table V reveal that the regression task is difficult to model accurately: even the best model achieves a MSE as high as 0.816, compared to an error of 1.0 achieved by predicting the mean of the output. Interestingly, the LARS ranking for variables results in the best accuracy, compared to around 0.85 for the DT, MI and RReliefF ranked variable sets.

F. Tecator

The Tecator data set consists of 240 samples of near infrared spectra with the task being to model the fat content of food products [45]. The spectrum is measured as 100 channels of different wavelengths, and consequently neighbouring channels are highly correlated.

The results of the feature selection and LS-SVM are presented in table VI. The most accurate model is obtained by the features selected by the Delta test. An interesting observation is that the 15 highest ranked variables by RReliefF correspond exactly to the range of channels 32–46. While all these variables are useful individually, selecting so many consecutive variables leads adds little new information, suggesting that the RReliefF method is unable to discriminate against features which provide only redundant information that is already contained in previously selected variables.

The value of the DT statistic itself in this case is exceptionally high, several times larger than the MSE of the resulting LS-SVM model. This implies that the 1-NN estimator is a rather poor prediction model, which is not surprising considering the high dimensionality and low number of samples. In spite of this, the method works well as a variable selection criterion, where only the relative accuracy matters.

TABLE V

THE SELECTED INPUTS AND MEDIAN LOO MSE FOR THE BREAST CANCER DATA. DT VALUE IS 0.4979.

		MSE
DT	2,3,9,12,15,21,32	0.8517
MI	1,2,3,7,8,9,10,11,12,14,19,23,24,25,26,27	0.8584
LARS 1–01	4	0.8968
LARS 1–02	4,31	0.8732
LARS 1–03	3,4,31	0.8342
LARS 1–04	3,4,30,31	0.8331
LARS 1–05	3,4,13,30,31	0.8255
LARS 1–06	3,4,11,13,30,31	0.8269
LARS 1–07	1,3,4,11,13,30,31	0.8261
LARS 1–08	1,3,4,10,11,13,30,31	0.8297
LARS 1–09	1,3,4,10,11,13,18,30,31	0.8304
LARS 1–10	1,3,4,10,11,13,18,28,30,31	0.8244
LARS 1–11	1,3,4,10,11,13,16,18,28,30,31	**0.8158**
LARS 1–12	1,3,4,10,11,13,16,18,28,30,31,32	0.8187
LARS 1–13	1,3,4,6,10,11,13,16,18,28,30,31,32	0.8221
LARS 1–14	1,3,4,6,10,11,13,16,18,19,28,30,31,32	0.8242
LARS 1–15	1,3,4,6,8,10,11,13,16,18,19,28,30,31,32	0.8206
LARS 1–16	1,3,4,6,7,8,10,11,13,16,18,19,28,30,31,32	0.8198
LARS 1–17	1,3,4,6,7,8,10,11,13,16,18,19,26,28,30,31,32	0.8252
LARS 1–18	1,3,4,6,7,8,10,11,13,16,18,19,20,26,28,30,31,32	0.8289
LARS 1–19	1,3,4,6,7,8,10,11,13,16,18,19,20,25,26,28,30,31,32	0.8336
LARS 1–20	1,3,4,6,7,8,10,11,13,16,18,19,20,22,25,26,28,30,31,32	0.8368
Relief 1–01	3	0.9424
Relief 1–02	3,26	0.9167
Relief 1–03	3,23,26	0.9221
Relief 1–04	3,23,26,30	0.9045
Relief 1–05	3,23,26,30,31	0.8888
Relief 1–06	3,23,26,27,30,31	0.8540
Relief 1–07	3,11,23,26,27,30,31	0.8601
Relief 1–08	3,7,11,23,26,27,30,31	0.8397
Relief 1–09	3,7,10,11,23,26,27,30,31	0.8443
Relief 1–10	3,7,10,11,23,26,27,28,30,31	0.8361
Relief 1–11	3,7,10,11,16,23,26,27,28,30,31	0.8449
Relief 1–12	3,7,10,11,16,22,23,26,27,28,30,31	0.8536
Relief 1–13	3,7,10,11,16,22,23,24,26,27,28,30,31	0.8526
Relief 1–14	2,3,7,10,11,16,22,23,24,26,27,28,30,31	0.8475
Relief 1–15	2,3,6,7,10,11,16,22,23,24,26,27,28,30,31	0.8501
Relief 1–16	2,3,4,6,7,10,11,16,22,23,24,26,27,28,30,31	0.8468
Relief 1–17	2,3,4,6,7,10,11,16,21,22,23,24,26,27,28,30,31	0.8531
Relief 1–18	2,3,4,6,7,10,11,16,21,22,23,24,26,27,28,29,30,31	0.8552
Relief 1–19	2,3,4,6,7,10,11,16,20,21,22,23,24,26,27,28,29,30,31	0.8467
Relief 1–20	2,3,4,6,7,10,11,16,18,20,21,22,23,24,26,27,28,29,30,31	0.8458
All	All 1–32	1.0052

TABLE VI

THE SELECTED INPUTS AND MEDIAN LOO MSE FOR THE TECATOR DATA. DT VALUE IS 0.0819.

		MSE
DT	5,12,16,38,39,40,41,42,49,50	**0.0122**
MI	7,40,41,42,43,48,50,53	0.0184
LARS 1–01	41	0.7137
LARS 1–02	7,41	0.0914
LARS 1–03	7,8,41	0.0882
LARS 1–04	7,8,41,63	0.0350
LARS 1–05	7,8,41,62,63	0.0440
LARS 1–06	7,8,41,56,62,63	0.0511
LARS 1–07	7,8,41,56,62,63,100	0.0226
LARS 1–08	7,8,41,56,59,62,63,100	0.0184
LARS 1–09	7,8,41,55,56,59,62,63,100	0.0357
LARS 1–10	7,8,41,55,56,59,62,63,64,100	0.0377
LARS 1–11	7,8,9,41,55,56,59,62,63,64,100	0.0320
LARS 1–12	7,8,9,41,54,55,56,59,62,63,64,100	0.0219
LARS 1–13	7,8,9,41,54,55,56,59,62,63,64,99,100	0.0159
LARS 1–14	7,8,9,41,53,54,55,56,59,62,63,64,99,100	0.0150
LARS 1–15	7,8,9,15,41,53,54,55,56,59,62,63,64,99,100	0.0148
LARS 1–16	7,8,9,15,41,42,53,54,55,56,59,62,63,64,99,100	0.0136
LARS 1–17	7,8,9,15,17,41,42,53,54,55,56,59,62,63,64,99,100	0.0139
LARS 1–18	5,7,8,9,15,17,41,42,53,54,55,56,59,62,63,64,99,100	0.0142
LARS 1–19	5,7,8,9,15,16,17,41,42,53,54,55,56,59,62,63,64,99,100	0.0166
LARS 1–20	5,7,8,9,15,16,17,41,42,52,53,54,55,56,59,62,63,64,99,100	0.0176
Relief 1–01	41	0.7137
Relief 1–02	40,41	0.6621
Relief 1–03	40,41,42	0.0500
Relief 1–04	39,40,41,42	0.0505
Relief 1–05	38,39,40,41,42	0.2482
Relief 1–06	38,39,40,41,42,43	0.0513
Relief 1–07	37,38,39,40,41,42,43	0.1844
Relief 1–08	37,38,39,40,41,42,43,44	0.1499
Relief 1–09	36,37,38,39,40,41,42,43,44	0.1486
Relief 1–10	35,36,37,38,39,40,41,42,43,44	0.1451
Relief 1–11	35,36,37,38,39,40,41,42,43,44,45	0.1218
Relief 1–12	34,35,36,37,38,39,40,41,42,43,44,45	0.1214
Relief 1–13	33,34,35,36,37,38,39,40,41,42,43,44,45	0.1205
Relief 1–14	32,33,34,35,36,37,38,39,40,41,42,43,44,45	0.1196
Relief 1–15	32,33,34,35,36,37,38,39,40,41,42,43,44,45,46	0.1044
Relief 1–16	9,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46	0.0141
Relief 1–17	9,10,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46	0.0144
Relief 1–18	8,9,10,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46	0.0147
Relief 1–19	8,9,10,11,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46	0.0150
Relief 1–20	7,8,9,10,11,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46	0.0156
All	All 1–100	1.0042

VII. CONCLUSIONS AND DISCUSSION

This paper presents the use of the "Delta test" 1-NN noise variance estimator for input variable selection. The theoretical analysis and experimental results support the notion that the method can provide desirable results in a wide variety of regression modelling problems. As the technique is simple it can be recommended as a suggested preprocessing step for nearly any regression task.

The theoretical claims presented in Section III involve minimising the expectation of the Delta test, which may seem insufficient considering that data often consists of a single realisation of some random process. However, as the variance becomes sufficiently small with a sufficient number of data points, using a single realisation is still effective.

For large problems, the computational cost of the method may become intractable with a naïve implementation. Hence care should be taken to appropriately implement both the evaluation of the nearest-neighbour search as well as how to explore the search space efficiently.

Further work regarding the use of the Delta test for variable selection involves exploring the precise extent of situations where the Delta test constitutes an appropriate method for variable selection. New, more efficient, search schemes for large data sets are also being developed. It is additionally of interest to study how the idea can be extended to other forms of dimensionality reduction, such as scaling or linear projection. Investigating the method's performance for variable selection in classification tasks is another appealing extension.

REFERENCES

[1] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, Eds., *Feature Extraction: Foundations and Applications*. Springer, 2006.

[2] C. J. Stone, *A Course in Probability and Statistics*. Duxbury Press, 1995.

[3] D. Evans and A. J. Jones, "Non-parametric estimation of residual moments and covariance," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, vol. 464, no. 2099, pp. 2831–2846, 2008.

[4] E. Eirola, E. Liitiäinen, A. Lendasse, F. Corona, and M. Verleysen, "Using the Delta test for variable selection," in *Proceedings of ESANN 2008, European Symposium on Artificial Neural Networks, Bruges (Belgium)*, Apr. 2008, pp. 25–30.

[5] F. M. Pouzols and A. B. Barros, "Automatic clustering-based identification of autoregressive fuzzy inference models for time series," *Neurocomputing*, vol. 73, no. 10–12, pp. 1937–1949, 2010, subspace Learning / Selected papers from the European Symposium on Time Series Prediction.

[6] S. Ben Taieb, A. Sorjamaa, and G. Bontempi, "Multiple-output modeling for multi-step-ahead time series forecasting," *Neurocomputing*, vol. 73, no. 10–12, pp. 1950–1957, 2010, subspace Learning / Selected papers from the European Symposium on Time Series Prediction.

[7] A. Sorjamaa, J. Hao, N. Reyhani, Y. Ji, and A. Lendasse, "Methodology for long-term prediction of time series," *Neurocomputing*, vol. 70, no. 16–18, pp. 2861–2869, Oct. 2007.

[8] F. Mateo and A. Lendasse, "A variable selection approach based on the delta test for extreme learning machine models," in *Proceedings of the European Symposium on Time Series Prediction*, Sep. 2008, pp. 57–66.

[9] Q. Yu, E. Séverin, and A. Lendasse, "A global methodology for variable selection: Application to financial modeling," in *Mashs 2007, Computational Methods for Modelling and learning in Social and Human Sciences, Brest (France)*, May 2007.

[10] F. Liébana-Cabanillas, R. Nogueras, L. Herrera, and A. Guillén, "Analysing user trust in electronic banking using data mining methods," *Expert Systems with Applications*, vol. 40, no. 14, pp. 5439–5447, 2013.

[11] R. Garcia-del Moral, A. Guillén, L. Herrera, A. Cañas, and I. Rojas, "Parametric and non-parametric feature selection for kidney transplants," in *Advances in Computational Intelligence*, ser. Lecture Notes in Computer Science, I. Rojas, G. Joya, and J. Cabestany, Eds. Springer Berlin Heidelberg, 2013, vol. 7903, pp. 72–79.

[12] Q. Yu, M. van Heeswijk, Y. Miche, R. Nian, B. He, E. Séverin, and A. Lendasse, "Ensemble delta test-extreme learning machine (DT-ELM) for regression," *Neurocomputing*, vol. 129, pp. 153–158, 2014.

[13] A. Guillén, D. Sovilj, F. Mateo, I. Rojas, and A. Lendasse, "Minimizing the Delta test for variable selection in regression problems," *International Journal of High Performance Systems Architecture*, vol. 1, no. 4, pp. 269–281, 2008.

[14] F. Mateo, D. Sovilj, R. Gadea, and A. Lendasse, "RCGA-S/RCGA-SP methods to minimize the delta test for regression tasks," in *Bio-Inspired Systems: Computational and Ambient Intelligence*, ser. Lecture Notes in Computer Science. Springer, 2009, vol. 5517, pp. 359–366.

[15] F. Mateo, D. Sovilj, and R. Gadea, "Approximate k-NN delta test minimization method using genetic algorithms: Application to time series," *Neurocomputing*, vol. 73, no. 10-12, pp. 2017–2029, 2010, subspace Learning / Selected papers from the European Symposium on Time Series Prediction.

[16] D. Sovilj, A. Sorjamaa, and Y. Miche, "Tabu search with delta test for time series prediction using OP-KNN," in *Proceedings of the European Symposium on Time Series Prediction*, Sep. 2008, pp. 187–196.

[17] D. Sovilj, "Multistart strategy using delta test for variable selection," in *ICANN 2011, Part II*, ser. Lecture Notes in Computer Science, T. Honkela, W. Duch, M. Girolami, and S. Kaski, Eds., vol. 6792. Springer, June 14–17 2011, pp. 413–420.

[18] A. Guillén, M. van Heeswijk, D. Sovilj, M. G. Arenas, L. J. Herrera, H. Pomares, and I. Rojas, "Variable selection in a GPU cluster using delta test," in *IWANN (1)*, 2011, pp. 393–400.

[19] A. Guillén, M. I. García Arenas, M. van Heeswijk, D. Sovilj, A. Lendasse, L. J. Herrera, H. Pomares, and I. Rojas, "Fast feature selection in a gpu cluster using the delta test," *Entropy*, vol. 16, no. 2, pp. 854–869, 2014.

[20] A. Navot, L. Shpigelman, N. Tishby, and E. Vaadia, "Nearest neighbor based feature selection for regression and its application to neural activity," in *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press, 2006, pp. 995–1002.

[21] D. François, *High-dimensional data analysis: from optimal metrics to feature selection*. VDM Verlag Dr. Muller, 2008.

[22] A. J. Jones, "New tools in non-linear modelling and prediction," *Computational Management Science*, vol. 1, no. 2, pp. 109–149, 2004.

[23] H. Pi and C. Peterson, "Finding the embedding dimension and variable dependencies in time series," *Neural Computation*, vol. 6, no. 3, pp. 509–520, 1994.

[24] Aðalbjörn Stefànson, N. Koncar, and A. J. Jones, "A note on the gamma test," *Neural Computing & Applications*, vol. 5, no. 3, pp. 131–133, 1997.

[25] J. Rice, "Bandwidth choice for nonparametric regression," *The Annals of Statistics*, vol. 12, no. 4, pp. 1215–1230, 1984.

[26] E. Liitiäinen, M. Verleysen, F. Corona, and A. Lendasse, "Residual variance estimation in machine learning," *Neurocomputing*, vol. 72, no. 16–18, pp. 3692–3703, 2009.

[27] M. D. Penrose, "Laws of large numbers in stochastic geometry with statistical applications," *Bernoulli*, vol. 13, no. 4, pp. 1124–1150, 2007.

[28] V. Spokoiny, "Variance estimation for high-dimensional regression models," *Journal of Multivariate Analysis*, vol. 82, no. 1, pp. 111–133, 2002.

[29] U. U. Müller, A. Schick, and W. Wefelmeyer, "Estimating the error distribution function in nonparametric regression with multivariate covariates," *Statistics & Probability Letters*, vol. 79, no. 7, pp. 957–964, 2009.

[30] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[31] F. Glover and M. Laguna, *Tabu Search*. Norwell, MA, USA: Kluwer Academic Publishers, 1997.

[32] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *Neural Networks, IEEE Transactions on*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[33] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, 2004.

[34] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[35] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, Oct. 2003.

[36] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.

[37] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the ninth international workshop on Machine learning*, ser. ML92. Morgan Kaufmann, 1992, pp. 249–256.

[38] G. C. Cawley and N. L. C. Talbot, "Fast exact leave-one-out cross-validation of sparse least-squares support vector machines," *Neural Networks*, vol. 17, no. 10, pp. 1467–1475, 2004.

[39] A. Asuncion and D. J. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/ mlearn/MLRepository.html

[40] A. S. Weigend and N. A. Gershenfeld, Eds., *Time Series Prediction: Forecasting the Future and Understanding the Past*. Reading, MA: Addison-Wesley, 1994.

[41] "The Santa Fe time series competition data," 1991. [Online]. Available: http://www-psych.stanford.edu/ andreas/Time-Series/SantaFe.html

[42] "AnthroKids — Anthropometric data of children," 1977. [Online]. Available: http://ovrt.nist.gov/projects/anthrokids/

[43] J. D. Hirst, R. D. King, and M. J. E. Sternberg, "Quantitative structure-activity relationships by neural networks and inductive logic programming. II. The inhibition of dihydrofolate reductase by triazines," *Journal of Computer-Aided Molecular Design*, vol. 8, pp. 421–432, 1994.

[44] L. Torgo, "Regression datasets," 2012, University of Porto. [Online]. Available: http://www.liaad.up.pt/ ltorgo/Regression/DataSets.html

[45] H. H. Thodberg, "Tecator data set," 1995. [Online]. Available: http://lib.stat.cmu.edu/datasets/tecator